LA-UR -89-1627

Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36

LA-UR--89-1627

DE89 012614

TITLE MULTIPLE CROSSBAR NETWORK: INTEGRATED SUPERCOMPUTING FRAMEWORK

AUTHOR(S) Randy L. Hoebelheinrich

SUBMITTED TO Supercomputing '89
Reno, Nevada
November 13-17, 1989

DISCLAIMER

Los Alamos National Laboratory
Los Alamos, New Mexico 87545

Multiple Crossbar Network: Integrated Supercomputing Framework

Randy Hoebelheinrich

Computing and Communications Division

Los Alamos National Laboratory

Los Alamos, New Mexico

ABSTRACT

At Los Alamos National Laboratory, site of one of the world's most powerful scientific supercomputing facilities, a prototype network for an environment that links supercomputers and workstations is being developed. Driven by a need to provide graphics data at movie rates across a network from a Cray supercomputer to a Sun scientific workstation, the network is called the Multiple Crossbar Network (MCN). It is intended to be a coarsely grained, loosely coupled, general-purpose interconnection network that will vastly increase the speed at which supercomputers communicate with each other in large networks. The components of the network are described, as well as work done in collaboration with vendors who are interested in providing commercial products.

INTRODUCTION

The world of supercomputing is in transition, and researchers at Los Alamos National Laboratory are placing greater demands on the computers than ever before. Because of their speed, supercomputers

are essential tools for doing science today, but these machines will become even more useful when we can connect them into networks so they can communicate with each other and with workstations as quickly as they process information.

Los Alamos National Laboratory has acted as the guiding influence in the development of a prototype high-speed network architecture called the Multiple Crossbar Network (MCN). The MCN, composed of high-performance switches and point-to-point channels, will vastly increase the speed at which supercomputers communicate with each other in large networks. Information is transmitted between computers through high-speed channels (HSCs). Another component of the MCN is the CrossPoint Star (CP*), a richly connected special-purpose switch designed to replace general-purpose computer packet switches at Los Alamos. The CP* switch is composed of special-purpose protocol processors, the HSC, and a crossbar switch (CBS) core. Services and protocols provided for the MCN include a data link protocol over the HSC with a channel access capability, intranet routing and network access protocols, and both a network management and name-server c pability. The current CP* has an aggregate bandwidth of 12 Gbit/s. Future versions are intended to have bandwidths of 24 and 48 Gbit/s.

The MCN will be a hierarchy of interconnection networks. The hosts themselves may include multistage interconnection networks (MINs) in their architecture. As an example consider Connection Machine's CM 2, Cray's X MP, BBN's Butterfly, or Intel's iPSC hypercube. The

transport network between hosts consisting of interconnecting CP*s will have a potentially rich connection scheme. Finally, the interconnection network within the CP* switch itself will use a crossbar interconnection network as opposed to a bus or ring based system. Addressing and routing in the MCN can then be flexible and consistent. This topology would also be potentially partitionable and reconfigurable. Distributed operating systems such as CMU's MACH or Apollo's DOMAIN could utilize this framework in association with an secure access control, auditing, and accounting scheme.

BACKGROUND

Los Alamos has been networking heterogeneous supercomputers for almost 15 years. The central network of the Integrated Computing Network (ICN) is based on a packet-switched network consisting of high-speed point-to-point channels and general-purpose computers for switching. Currently, the packet switches are Concept 32/67s, manufactured by Gould, Inc. Current network channels, designed at Los Alamos, transmit information at up to 50 Mbit/s. The protocols were also designed and developed at Los Alamos. The central protocol is a datagram service with additional security features. This architecture has served the Laboratory well. However, limitations in the current network and new requirements and applications have motivated efforts to come up with a new solution to supercomputer networking at Los Alamos. The computation speed of supercomputers has outpaced the I/O bandwidth of their channels. New supercomputer I/O channels are being developed that attempt to keep

up with the computational speed. New networking solutions will need to utilize these faster channels.

Concurrently, progress is being made in parallel computation and visualization of data. As more people use workstations in a distributed environment, the network must be even more reliable than our current network capability. A more complex network also indicates a need for better management, better fault handling, and an integrated structure of servers and applications. We obviously need to move toward distributed, parallel, and high-speed computing. The areas of hardware changes centered on the switches and channels to them. We needed a high-speed channel, switching at the physical layer, and a special-purpose protocol processor on each channel. In addition we learned from years of experience that a standard interface to the network was essential. It was also necessary to design the means to allow detailed control and management of the network transport hardware and software.

CHRONOLOGY

The sequence of events for this project are given below:

April 1986	Ultra-High-Speed Graphics Project
July 1986	Initial HSC Standard Proposed by LANL
July 1986	CP* Concept Proposed
January 1987	MCN Concept Developed
December 1987	LANL/DEC CP* Collaboration

January-July 1988	Service and Protocol Specification
January 1988	Crossbar Interface (CBI) Design Initiated
May 1988	Initial HSC Data Link Proposed by LANL
July 1988	CBS Project Initiated
November 1988	Fiber HSC Standard Initiated
January 1989	CBS Assembly
February 1989	HSC Interface for CBI
February 1989	CBI Delivery
May 1989	HSC Public Review
March-May 1989	Base Level 0 CBI-CBS Testing
May-June 1989	Base Level 1 Data Link Testing
June-September 1989	Base Level 2-5 Intranet, Network Access Testing

MOTIVATION

Motivation for this new architecture came from many areas. The primary areas were a need for increased performance and reliability and better network management. These needs are nothing new. There was, however, an interesting combination of parallel events and activities that culminated in the initiation of efforts to solve these supercomputer networking requirements at Los Alamos. These activities could be summarized as efforts to explore, realize, and utilize the power of parallelism, distribution, and visualization. In concrete terms, work in interconnection networks, complex systems, and simulations needed more computing power in the form of multiple computer solutions, increased network bandwidth, and finally, lower

latency across the network for remote procedures. Previous ICN switches were general-purpose computers running protocol processing and switching code. Each of these nodes serviced a number of high-speed lines (50 Mbit/s). The configuration included a central controller, a serially accessed bus, central memory, and multiple I/O devices. The bandwidth from the supercomputers to the users' workstations would eventually outpace the capability of this system. A solution was necessary.

Our need and motivation began with a simple requirement to provide graphics data at movie rates across the network from a Cray supercomputer to a frame buffer display. This meant a high-speed channel and some sort of switching system. Because of security requirements, the switching network couldn't be a shared broadcast network. Because we wanted general-purpose networking and no blocking of resources (such as one Cray channel dedicated to one user), we needed packet switching vs. circuit switching. The data transmission requirements from a Cray to a frame buffer are 10-30 frames per second. This performance is needed to support visualization. If this were to be generalized, it would mean more would be needed than a pipe or high-bandwidth dedicated channel. A network that could support this kind of bandwidth to a user's desktop workstation was required. Since LANL needs to maintain the privacy and generalized or flexible topology of a packet switched network, the previous switched model was retained.

Some people thought that one supercomputer working on a problem or simulation was simply not enough. Interconnection networks were becoming more common in supercomputer and multiprocessor systems. These systems are tightly coupled and fine-grained. There was an increasing use for more loosely coupled and course-grained systems as well.

SOLUTION

Several design characteristics were viewed as important for a new, high-bandwidth switch. Most important was distributing the processing overhead and minimizing decision overhead at the switch junction. This could be accomplished by having a processor memory unit dedicated to each channel or to each bidirectional set of channels. At that point, the units could talk to each other by a physical data transport media. Rather than use a serial shared-bus system, a CBS was used. In this way, arbitration was confined to those channels needing access to a given destination, and contention was limited to a control processor in the interconnection network. This crossbar interconnection network needed to exist on some sort of channel between source and destination switches or nodes. Los Alamos already had the High-Speed Parallel Interface (HSPI). This parallel concept could be extended and revised to be higher speed and wider data path with a different connect scheme incorporating a multiple access mechanism for channels. In addition, a channel was needed to connect commercial systems to this network. It was desireable to have a standard specification channel for everyone. Finally, a

transparent physical switching capability was needed on this channel. This final feature would effectively distribute actual switching functionality away from the protocol processors and out "on the wire." The result was CP*.

CP*

The CP* can be thought of as a packet-switching node. CP* is made up of three major elements: (1) the HSC, (2) the CBS, and (3) the CBI. CP* is designed to increase performance by having distributed special-purpose protocol processing on each channel and incorporating physical layer switching between these processors. The physical layer switching is accomplished on an HSC with the aid of an intermediate CBS controller. This CBS is strictly dedicated to switching links to minimize switching latency and provide fast packet switching. The distributed protocol processors or CBIs provide streamlined optimal packet throughput. By using this design, we will, in a broader sense, have distributed the functions of the traditional packet-switching node over many processors and controllers at the channel end points as well as "on the wire." This is replicated on all links for parallel simultaneous transfers at any CP* node.

HSC

There were a number of goals for the HSC. First, it had to be high speed. An 800-Mbit/s channel already existed on the Cray computers, so matching that data rate seemed appropriate. Second, we knew

from experience that standardizing the HSC interface for vendor implementations was highly desireable. The question was, would industry see sufficient need for such a high-speed point-to-point channel? History shows there was considerable interest. Finally, and this is the most interesting goal, physical layer switching by means of an intermediate CBS and controller on the HSC was desireable. This would prove crucial in the overall architecture of LANL's new supercomputer network.

CBS

The CBS's sole purpose is connecting a Source HSC to a requested Destination HSC. This request is made by a single parameter supplied by the Source HSC at connect time.

CBI

The CBI, which has been designed and built by Digital Equipment Corporation (DEC), is a protocol processor for CP*. It manages network packets for two HSC sources and two HSC destinations. The CBI's primary purpose is to act as a specialized processor for protocols. This allows lowering of protocol processing overhead and offloading protocol processing from the hosts using the network. Use of a RISC processor, VRAM, hardware FIFOs and limiting interfaces to four channels also streamlines packet processing.

CP* as a network switch is a modular component of a network. By combining any number of CP*s in any connection scheme, an MCN results. Software protocols were necessary. Los Alamos designed an initial HSC physical layer protocol, a data link protocol to control the HSC, an intranet routing protocol, a network access protocol, and a name-server service. The intranet protocol was implemented by the CBIs alone, while the network access protocol is to be implemented by vendors wishing to access the MCN.

SERVICES AND PROTOCOLS

Services and protocols provided for the MCN include a data link protocol over the HSC with a channel access capability, intranet routing and network access protocols, and both a network management and name-server capability.

The link configurations in the MCN are of two types. One is a point-to-point simplex HSC. The second is a point-to-point simplex HSC with an intermediate CBS between the source and destination HSC entities. It is helpful to view this latter configuration from the perspective of three different elements. These elements are the HSC, the source data link entity, and the intermediate crossbar switching core. The link to the HSC is point-to-point where the switching core is transparent. The data link entity will view the link as a multipoint configuration with physical switching between several HSCs. The crossbar switching core will also view the set of HSCs as a multipoint topology. In all cases, the link duplexity is simplex.

DATA LINK

The data link for the MCN uses channel access control as a form of link selection. The data link will use physical address to HSC mapping and the HSC connection sequence as a channel access control mechanism. This is accomplished by utilizing the underlying HSC, CBS, and I-Field to access one of many possible destination HSCs. The data link entity should be viewed as contending for an available HSC in a multipoint configuration of data link entities. The data link entity does not control multiple HSCs simultaneously. It does not, therefore, provide a downward multiplexing or splitting capability.

INTRANET

The intranet accepts packets from a network access entity and transfers these packets over a series of HSC links to a destination network access entity. The intranet is a local connectionless data transfer service. Its primary function is to take each packet and determine the routing for the packet.

The intranet receives a destination physical network address or SNPA from the source network access entity when it is given a packet to transmit. This address is used for link selection over the CBS.

Security and access information are checked after entry to and before exit from the MCN.

NETWORK ACCESS

The network access accepts packets from a transport entity and transfers these packets through a data link to a destination network access entity via the intranet. Its primary function is to transfer data to a logical network destination or NSAP. The network access receives a network name when it is given a packet to transmit. The network access provides a server function that will translate this name to a logical network address or NSAP to be used by the network access sublayer for transferring the packet. This is fully defined in the name-server documentation.

The network access protocol also takes each packet and acts as a security, authorization, and flow control boundary to the MCN.

NAME-SERVER SERVICE

The primary functions of the name-server service are twofold. First, it is used to hold a service to logical address translation database for name translations on the network access portion of a host and logical-to-physical address translations for the intranet protocol. The latter capability is designed to allow flexibility for configurations and system movement by adding a level of indirection to the addresses. Second, the name server service will allow introduction, maintenance, access control, and accountability of objects that are or become part of the

network. These objects include, hosts, processes, users, and user sessions, to name a few.

IMPLEMENTATION

What has been accomplished at this point? HSC or HSC-like implementations have been completed and, in some cases, announced as products by various commercial interests.

DEC has built four CBIs and installed them at Los Alamos where they are undergoing testing with the CBS and Los Alamos HSC. The first level included passing data without checking by the RISC processor and HSC interfaces on the CBI. The CBS has also been included in the loop and demonstrated to work.

At progressive stages, more protocols will be added to the CBIs up to and including the network management services that reside above the intranet and network access protocols. This work will also include implementations of the HSC, LANL's original data link, and network access protocols on certain vendor equipment that we have agreed to include in an MCN testbed. This equipment will include supercomputers, workstations, and frame buffers. All of this equipment will access the MCN with HSCs.

FUTURE

Efforts in the future will include connecting a full set of hosts to the MCN and developing a full range of tests and applications to verify and improve the protocols and services at all levels. In particular, the network access connections in both hardware and software need to be investigated to begin reducing data copy, share network address space with users, and verify security and access control. We are also interested in developing some form of capability based routing for partitioning and need-to-know computing. Ultimately, the MCN can act as a network-centered transport system. One can view the MCN as a collective computer with hosts effectively used as peripherals to an overall system controlled by a distributed operating system. In addition, it will be advantageous to think of the host front-end to the MCN as a peer coprocessor to the host itself. In this way, the host front-end processor handles all distributed operating system, session, and communication tasks, while the other processor handles the computation or specific task it was designed to do.

CONCLUSION

We are now working with a prototype CP* that will soon be connected to three supercomputers, a frame buffer, and two workstations. All of these systems are from different vendors. With a proposed standard network access channel, vendors have some hope of utilizing the testbed. We have transferred data through an initial CP* configuration and will soon be expanding this setup. We expect to make

considerable progress between now and July 1989, when we plan to have CBIs with a full set of protocols communicating to attached hosts.

ACKNOWLEDGMENTS

I would like to acknowledge the efforts of Michael McGowen for the basic concepts for the HSC and CP*, which grew out of efforts on the Ultra-High-Speed Graphics Project at Los Alamos. Valuable contributions were also made by Richard Thomsen, Don Tolmie, Gene Domhoff, Steve Tenbrink, John Morrison, Dave Dubois, Allan Meddles, Eric Vandevere. We are also indebted to Norm Morse for supporting this project and Karl-Heinz Winkler for demonstrating the need. We are grateful to the members of DEC's Southwest Engineering group in Albuquerque who contributed greatly to this effort. Finally, credit goes to all companies and attendees of the HSC Working Group for their work in developing the HSC standard. This has truly turned into an industry-wide effort.

REFERENCES

(We still need to fix these up before sending off the paper.)

1. Los Alamos Network Architecture: Network Hardware

2. Los Alamos Network Architecture: Network Software and Protocols

3. High Speed Channel (HSC) Mechanical, Electrical and Signalling Protocol Requirements

4. Los Alamos Network Architecture: Nameserver Software and Protocols

5. Los Alamos Network Architecture: Data Link Service and Protocol Specifications

6. Los Alamos Network Architecture: Network Access Service and Protocol Specifications

7. Los Alamos Network Architecture: Intranet Service and Protocol Specifications

8. Los Alamos Network Architecture: Network Management Service and Protocol Specifications

9. Los Alamos Network Packet Switching Node Security Functions